PREDICTION OF THE MOST EFFECTIVE ADJUVANT THERAPEUTIC COMBINATIONS FOR BREAST CANCER PATIENTS USING MULTINOMIAL CLASSIFICATION

MEROUANE ERTEL[1],AZEDDINE SADQUI[2], SAID AMALI[3],NOUR-EDDINE EL FADDOULI[4]

[1,2]Informatics and Applications Laboratory (IA), Faculty of Sciences, Moulay Ismail University, Morocco

[3]Informatics and Applications Laboratory (IA), FSJES, Moulay Ismail University, Morocco

[4]RIME Team, MASI Laboratory, E3S Research center EMI, Mohammed V University, Morocco

E-mail: [1] m.ertel@edu.umi.ac.ma,[2]a.sadqui@umi.ac.ma,
[3]s_amali@yahoo.com,[4]noureddine.elfaddouli@um5.ac.ma

ABSTRACT

The main goal of precision medicine in the fight against cancer is to predict effective treatment modalities based on the unique molecular genetic profiles of a tumor. Understanding the factors that influence treatment success is critical because people with breast cancer at similar stages respond differently to treatment. In order to reduce the likelihood of recurrence of metastases in breast cancer patients, this study proposes a supervised multinomial logistic regression model. This model will help clinicians make decisions about which treatment plans they should recommend to patients. In addition, this article compares a number of polynomial machine learning technologies, including Naive Bayes, Decision Tree, Support Vector Machine, Random Forest, and Neural Network (ANN). Accuracy results for adjuvant treatment combination prediction show that the Random Forest classifier is more accurate.

Keywords: *Machine Learning; Multinomial Logistic Regression; Personalized Medicine; Multi-Class Classification; Adjuvant Therapy; Breast Cancer.*

1. INTRODUCTION

Breast cancer is the second most common cancer and the most prevalent malignancy among Moroccan women [1]. Despite therapeutic progress, better treatment outcome prediction is still required in order to choose the most efficient treatment plan and prevent the growth of metastases and the return of the original tumor. For newly diagnosed cases of breast cancer, surgery (of various kinds) is the primary line of treatment, followed by adjuvant therapy (chemotherapy, radiation, hormone therapy, targeted therapies, and their mixes) [2]–[4]. Once the tumor has been removed, it is essential that the attending physician choose an adjuvant therapy capable of eliminating small pockets of malignant cells which, if left untreated, could grow and become metastatic [5]–[7].

To identify the most effective therapy approach and enhance patient outcomes, a personalized strategy over time is required. Machine learning algorithms have been widely employed in personalized medicine, particularly in oncology, where numerous techniques for predicting therapy response based on data have evolved (clinical, pathological, and biological) [8]–[12].

To do this, doctors choose a combination of treatments (chemotherapy, hormone therapy, radiation therapy, targeted therapy) with the goal of achieving the best possible outcome for a patient. The chosen treatment depends on the characteristics (clinical, histological, molecular). Currently validated chemotherapies mainly include anthracyclines and taxanes[13]. The use of trastuzumab, a monoclonal antibody targeting HER2, in combination with chemotherapy, is systematic in the event of overexpression or amplification of HER2[14]. When the tumor expresses estrogen (ER) and/or progesterone (PR) hormone receptors, there is an indication for adjuvant hormone therapy in premenopausal patients, sequential treatment: 2 years of tamoxifen then 3 years of aromatase inhibitors (IA) or 2 years are recommended to reduce the risk of recurrence[15]. Individual patient responses to a given treatment can differ significantly. Although a treatment regimen works well for one patient, it may not work for another. Identifying the optimal treatment for an individual patient, from the complex array of options, can be incredibly difficult.

Over the years, identifying the most effective treatment regimen and improving patient outcomes requires a tailored approach. Many machine learning algorithms have been used in personalized medicine, especially in the field of oncology, where many methods for predicting treatment response have emerged, based on (clinical – pathological – biological) data for the treatment response. Identification of subgroups of patients with early breast cancer. Recently, some researchers have been working to develop web-based models that can take into account a large amount of data from cancer registries, to help determine the need adjuvant therapy in patients with early stage breast cancer[16]–[18].

The paper uses various machine learning techniques including Logistic Regression (LR), Naive Bayes (NB), Random Forest (RF), Decision Tree (DT), Support Vector Machine (SVM) and Artificial Neural Network (ANN). These techniques apply to all the data collected at the regional oncology center in Meknes-Morocco.

The main contributions of this study are:

- ✓ A supervised multinomial logistic regression model to predict the combination of effective therapies that reduces the risk of metastatic relapse in breast cancer patients based on machine learning.
- ✓ Identification of effective variables in learning our model.
- ✓ Use coding techniques to code adjuvant treatments that represent the target variable.

The rest of the paper is organized as follows. Section 2 discusses the relevant work involved in this Region. Section 3 includes materials and methods used. Section 4 discusses the results and finally Section 5 concludes the article.

2. RELATED WORKS

Many studies have been published in the literature describing data-driven actionable intelligence to guide the use of adjuvant therapies for breast cancer.

Several machine learning algorithms have been created to extract knowledge from databases, including supervised learning techniques. These algorithms are most often used to predict optimal treatments by maximizing relapse-free survival for breast cancer patients. This section summarizes various studies relevant to this problem.

The PREDICT tool is one of them; it uses multivariate statistical analysis to estimate a person's likelihood of surviving based on the fusion of clinical factors. Using this tool to plan adjuvant treatment is strongly recommended [19]. Other tools, such as the Adjutorium, used to determine if patients need adjuvant therapies (chemotherapy and hormone therapy) provided in addition to surgery. In 2014, IBM released Watson for Oncology in order to employ machine learning to advise a cancer treatment regimen [20]. In line with these studies and the developments of predictive machine learning models that allow clinicians to accurately predict the effectiveness of each adjuvant therapy combination, we find it useful to conduct this study, with the addition of new variables representing drugs used in cancer chemotherapy, hormone therapy and targeted therapies, which can give a breakdown of how each treatment has contributed to relapse-free survival rates and which will help us predict which treatment combination will be effective in treating a breast cancer patient.

3. MATERIALS AND METHODS

3.1 Data Understanding

3.1.1 Data Source

This predictive study included breast cancer patients localized to all histological types of breast cancer in the Regional Oncology Center of Meknes in Morocco, during the period 2014 to 2021, who underwent surgery associated with adjuvant therapy during the years 2014 - 2016 (chemotherapy - targeted therapies - radiotherapy - hormonal therapy) with a follow-up of at least 60 months.

The dataset for our system consists of 511 entries and 15 variables. These variables, including the target variable, give demographic, clinical, and therapeutic information about the patient (adjuvant treatments). The information was gathered from a database of patient records, and it was verified by professionals (treating physicians).

3.1.2 Dataset features

Characteristics of tumor variables, patient follow-up and treatment outcome were recorded in the system by the treating physicians. The following information was extracted from each patient: age at diagnosis of breast cancer, menopause, size of the primary tumor (TS), histological grade of the tumor, number of axillary lymph nodes involved, cell marker of proliferation (Ki67), estrogen and progesterone receptor expression (HR: Negative (0) / Positive (1), type of surgery (lumpectomy (1) / mastectomy (2)), epidermal growth factor receptor-2 (Her2: negative/positive), as well as the variables of the adjuvant treatment protocols (Type of chemotherapy, Hormone therapy, Herceptin and radiotherapy), Table 1 summarizes the main variables of our study.

Table 1: Demographic And Cancer-Specific Information Of Patients And Treatment

Variable_Name		Definition
Age_diagnosis		20-34 ; 35 to 44 ;45-54 ; 55 ≥ years old
Postmenopausal		0 = Before the age of 50 1 = age of 50
Tumor_size		≤2 ; 3-4 ; ≥5
Lymph_nodes		0="no";1="1–3";2="4–9"; 3 = " >9"
Tumour_grade		1 ; 2 ; 3
Her2		0 = "Negative" ; 1="Positive"
ER		0 = "Negative" ; 1="Positive"
PR		0 = "negative" ;1 = "positive"
Ki67		≤ 14 % ; 15-24 % ; 25 -29 % ; ≥30
Surgery_type		1 = " Lumpectomy " 2 = " Mastectomy "
Chimiotherapy	Anthracyclines	"NO"; "AC60"; "EC50"; "EC 100”
	Taxane	"NO"; "PACLITAXEL"; "DOCETAXEL"
Herceptine		"NO"; "TRASTUZUMAB"
Radiotherapy		"NO"; "YES"
Hormonotherapy		"NO"; "TAMOXIFENE"; "LETROZOLE"; "ANASTROZOLE"; "EXEMESTANE";"TAMOXIFENE + AROMATASE "; "INHIBITORS(AIS)";"AROMATASE";"INHIBITORS (AIS) + TAMOXIFEN"

3.2 Multinomial logistic regression model

One of the key techniques utilized when the dependent variable is a nominal with more than two levels is the multinomial logistic regression model. Similar to multiple linear regressions, multinomial regression is a type of predictive analysis. Multinomial regression is used to describe the relationship between a nominal dependent variable and one or more independent variables [8], [21], [22] . After converting dependency into a logit variable, which is the natural logarithm of the likelihood that the dependant is equal to or not equal to some value, multinomial logistic regression employs maximum true semblance estimation (usually 1 in binary logistic models, or the highest value in multinomial models)[23][24]. Using logistic regression, one may calculate the likelihood that a certain event (or value) will occur. This indicates that, unlike ordinary least squares (OLS) regression, logistic regression evaluates changes in the dependant's log probability rather than the dependent itself [25].

In this study, we present a multinomial classification model that is based on clinico-biological information, industry standards, and recommendations for adjuvant therapy. In order to create predicted outcomes, we developed our predictive model based on the four most popular adjuvant treatment strategies, each of which has five treatment options (Anthracyclines, Taxanes, Herceptin, Radiotherapy, and Hormonotherapy) (Figure 1). We studied the adjuvant treatment strategies used in the regional oncology center of Meknes [26]. This model forecasts treatment regimens that successfully lower the chance of metastatic recurrence in breast cancer patients.

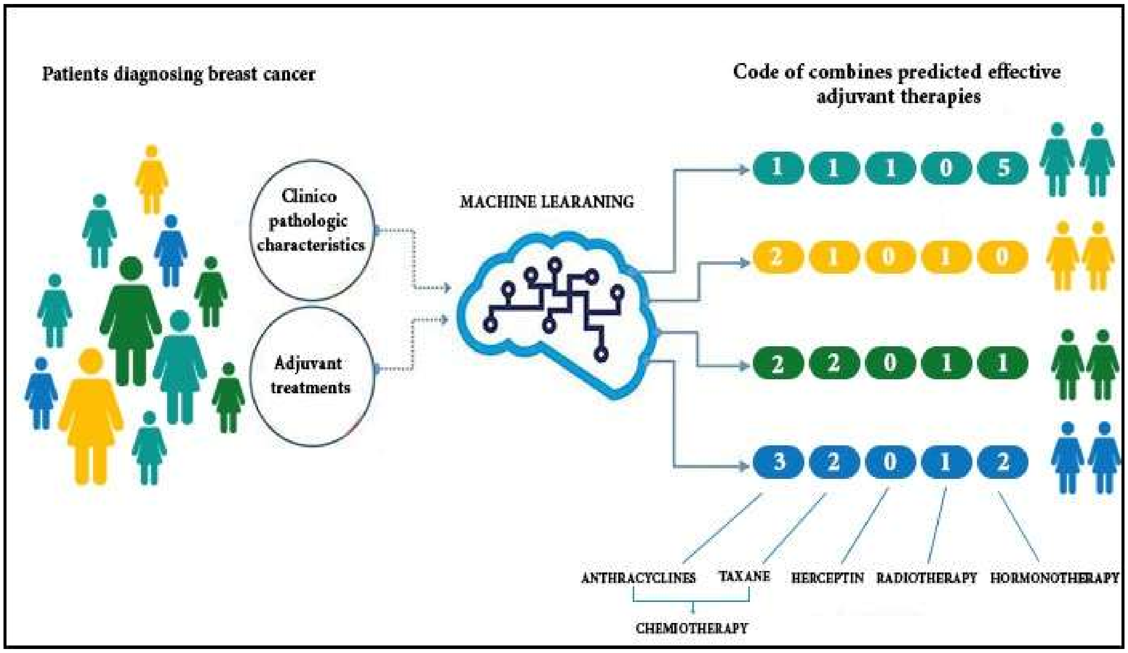


Fig.1 Our Predictive Model Of Adjuvant Treatment Combinations

In the part that follows, we will go through all of the data preparation steps that went into creating our study's multinomial prediction model.

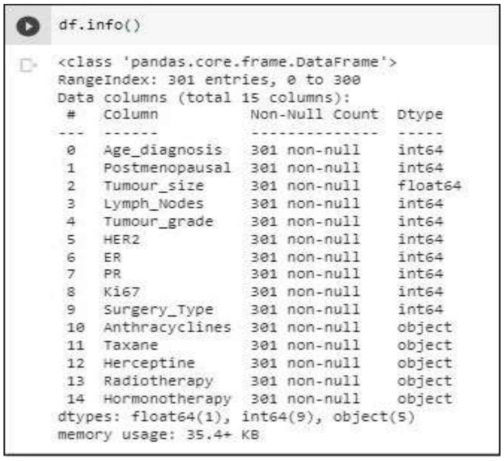
3.3 Data Preparation

3.3.1 Data cleaning

511 women who underwent adjuvant therapy for breast cancer were included in our study and included in our system database. The database has undergone a cleaning process in order to eliminate and reduce noise. We excluded 148 individuals whose tumors recurred throughout the study period, and discarded 62 registries with incomplete data. Finally, a data set of 301 records (see Table 2) was produced, each representing a unique case of breast cancer treated with a specific adjuvant treatment regimen.

In Table 2, each case can be represented by 15 distinct attributes: attributes between 0 and 8 represent different clinico-pathological characteristics, attribute 9 represents the type of surgery (independent variables), and attributes between 10 and 14 represent the type of adjuvant treatments used, i.e., dependent/categorical variables.

Table 2 :Dataframe Information



```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 301 entries, 0 to 300
Data columns (total 15 columns):
 #   Column           Non-Null Count  Dtype  
---  ------           --------------  -----  
 0   Age_diagnosis    301 non-null    int64  
 1   Postmenopausal   301 non-null    int64  
 2   Tumour_size      301 non-null    float64
 3   Lymph_Nodes      301 non-null    int64  
 4   Tumour_grade     301 non-null    int64  
 5   HER2             301 non-null    int64  
 6   ER               301 non-null    int64  
 7   PR               301 non-null    int64  
 8   Ki67             301 non-null    int64  
 9   Surgery_Type     301 non-null    int64  
 10  Anthracyclines   301 non-null    object 
 11  Taxane           301 non-null    object 
 12  Herceptine       301 non-null    object 
 13  Radiotherapy     301 non-null    object 
 14  Hormonotherapy   301 non-null    object 
dtypes: float64(1), int64(9), object(5)
memory usage: 35.4+ KB
```

In Table 2, each case can be represented by 15 distinct attributes: attributes between 0 and 8 represent different clinico-pathological characteristics, attribute 9 represents the type of surgery (independent variables), and attributes between 10 and 14 represent the type of adjuvant treatments used, i.e., dependent/categorical variables.

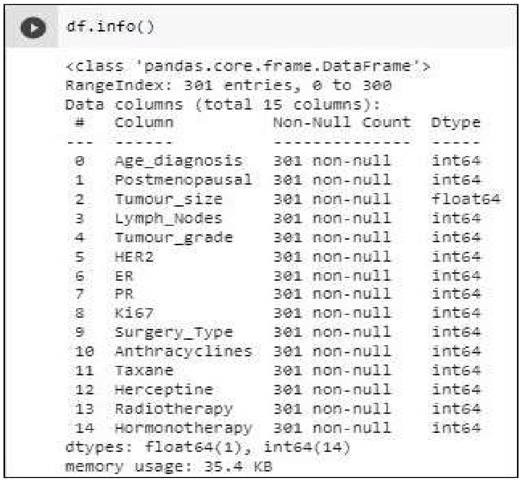
3.3.2 Label Encoding

Many machine learning algorithms work best when the variables are numeric. For this purpose, we used the function "convert_objects (convert_numeric=True) in the python library to convert the target values of the object data type to numeric (Anthracyclines - Taxane -Herceptin - Radiotherapy - Hormone therapy) (see Table 3) . The output displays the list of variables/features (see Table 4).

Table 3: Method for encoding the four adjuvant treatment strategies used in this study

				Coding	Treatment_Protocols_code
Chemotherapy	Anthracyclines	Epirubicine	No	0	0 or 1 or 2 or 3
			EC 50	1	
			FEC 100	2	
		Doxorubicine	AC60	3	
	Taxane	No		0	0 or 1 or 2
		Paclitaxel		1	
		Docetaxel		2	
Herceptin	No			0	0 or 1
	Trastuzumab			1	
Radiotherapy	No			0	0 or 1
	Yes			1	
Hormonotherapy	No			0	0 or 1 or 2 or 3 or 4 or 5 or 6
	Tamoxifene			*1*	
	Letrozole			*2*	
	Anastrozole			*3*	
	Exemestane			*4*	
	Tamoxifene + Aromatase Inhibitors (ai)			*5*	
	Aromatase inhibitors (ai) + Tamoxifene			*6*	

Table 4:Dataframe information after encoding



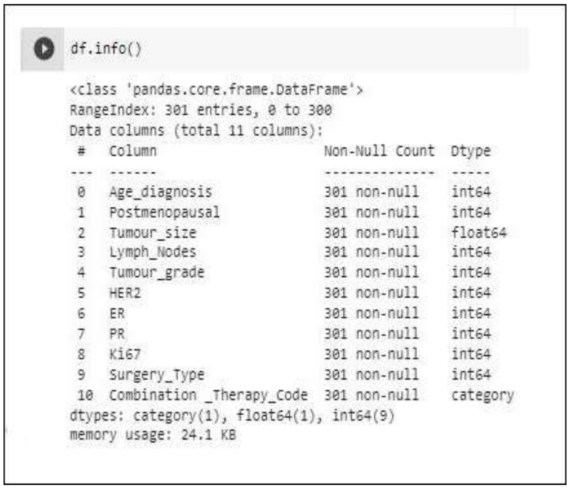
```
df.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 301 entries, 0 to 300
Data columns (total 15 columns):
 #   Column           Non-Null Count  Dtype
---  ------           --------------  -----
 0   Age_diagnosis    301 non-null    int64
 1   Postmenopausal   301 non-null    int64
 2   Tumour_size      301 non-null    float64
 3   Lymph_Nodes      301 non-null    int64
 4   Tumour_grade     301 non-null    int64
 5   HER2             301 non-null    int64
 6   ER               301 non-null    int64
 7   PR               301 non-null    int64
 8   Ki67             301 non-null    int64
 9   Surgery_Type     301 non-null    int64
 10  Anthracyclines   301 non-null    int64
 11  Taxane           301 non-null    int64
 12  Herceptine       301 non-null    int64
 13  Radiotherapy     301 non-null    int64
 14  Hormonotherapy   301 non-null    int64
dtypes: float64(1), int64(14)
memory usage: 35.4 KB
```

3.3.3 Conversion to Categorical Data:

In this study, we created a new numeric column using the concatenation technique in the Panda python library, which combines the encoded variables representing the adjuvant treatment variables (Anthracyclines - Taxane - Herceptin - Radiotherapy - Hormone therapy) in sequence. This new variable is named "Combination_Therapy_Code". Then, we converted this target variable "Combination_Therapy_Code" into a categorical variable composed of 05 values , column N°10 (see Table 5), to build a multi-class predictive model.

Table 5:Dataframe information after encoding



```
df.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 301 entries, 0 to 300
Data columns (total 11 columns):
 #   Column                    Non-Null Count  Dtype
---  ------                    --------------  -----
 0   Age_diagnosis             301 non-null    int64
 1   Postmenopausal            301 non-null    int64
 2   Tumour_size               301 non-null    float64
 3   Lymph_Nodes               301 non-null    int64
 4   Tumour_grade              301 non-null    int64
 5   HER2                      301 non-null    int64
 6   ER                        301 non-null    int64
 7   PR                        301 non-null    int64
 8   Ki67                      301 non-null    int64
 9   Surgery_Type              301 non-null    int64
 10  Combination _Therapy_Code 301 non-null    category
dtypes: category(1), float64(1), int64(9)
memory usage: 24.1 KB
```

According to Table 5, we present a new dependent variable that combines adjuvant treatments for women with breast cancer. Thus, the predictive outcome would be a combination of five-digit categories, each representing a different type of protocol (Anthracyclines-Taxane -Herceptin - Radiotherapy - Hormone therapy).

3.3.4 Correlation features

After the data transformation step, we need to check the relationship between the variables used in our study to develop a reasonable prognosis. Thus, a correlation matrix was performed using the Seaborn library to investigate the relationship between the features in the data set (see Figure.2).

Next, we calculate the correlation between all predictors and target responses, as shown in Figure. 2. A high correlation implies that there is a relationship between the independent variables and the target variable.

In this study, we included variables with high correlation because they have the highest predictive power (signal), and leave out variables with low correlation because they are probably less relevant. Although including more relevant features during training helps to improve predictive power, we always include all features in model training and then gradually exclude irrelevant features because it is not always possible to know in advance which features have a strong predictive influence.

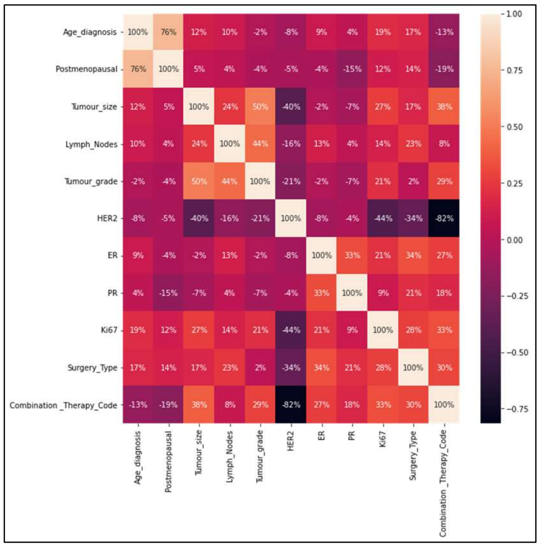


Fig.2 Heat Map For Checking Correlated Columns For Breast

3.4 Modiling

The modeling procedure, which involves teaching the machine learning algorithms to predict the classes, comes after the data pretreatment stage. We applied the following machine learning techniques in this study: Support Vector Machine (SVM), Naive Bayes (NB), Decision Tree (DT), Random Forest (RF), Logistic Regression (LR), Naive Bayes (NB), and Artificial Neural Network (ANN). The most popular algorithms for this type of categorization issue are these ones. Based on baseline demographics and clinical features, these classification models were employed in this study to classify successful individual adjuvant medication combinations in early breast cancer patients that minimize metastatic relapse. We employed the Python scikit-learn package to examine the data. Figure 3 depicts the experiment's entire procedure.

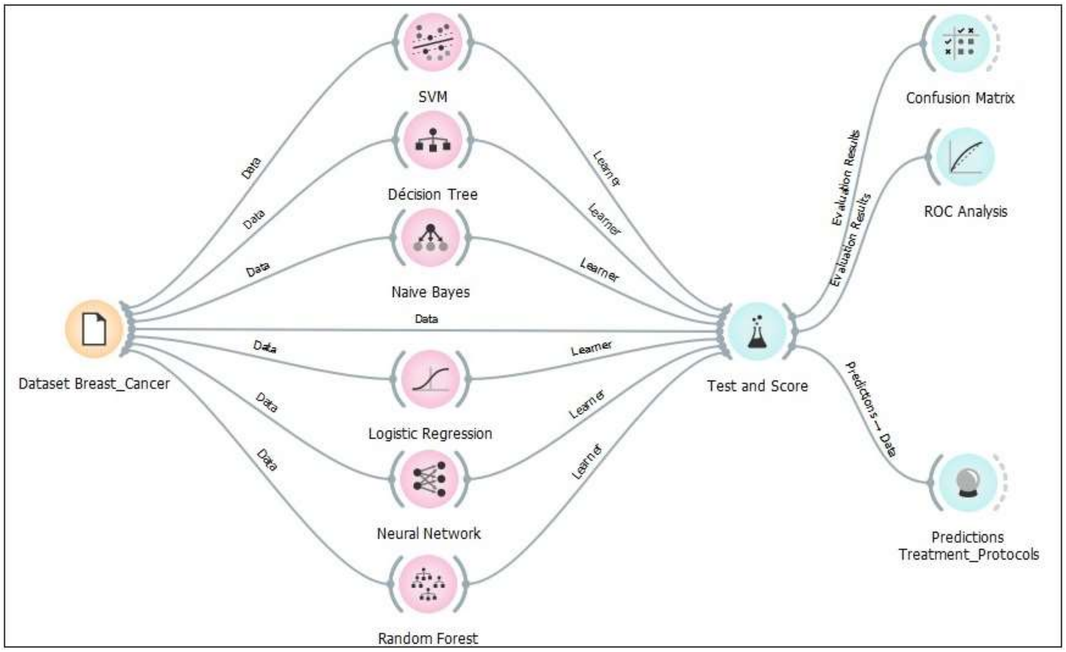


Fig.3 Model Machine Learning Use

To evaluate the classifiers, we applied a 10-fold cross-validation test which is a technique used to evaluate predictive models that partitions the original set into a training sample to train the model, and a set of tests to assess it in terms of effectiveness and efficiency.

3.4.1 Machine Learning Algorithms

3.4.1.1 Support-vector machines (SVM):

SVM is a machine learning technique based on statistical learning theory, used for classification and regression. SVM provides a better classification which generates a more complex boundary between classes [27]. SVM was chosen as one of the learning techniques to test the performance of the model because it better captured the fundamental properties of the data despite its small size[28]–[30].

3.4.1.2 Decision Tree:

The data is split multiple times into decision tree models based on the feature cutoff parameters. Different subsets of the dataset are formed as a result of the split, with each instance belonging to one of them. The end nodes or leaves are the ultimate subsets, while the inner nodes or split nodes are the intermediate subsets. The average score of the training data in this node is used to predict the score in each leaf node. Classification and regression can be performed with decision trees [31].

3.4.1.3 Naïve Bayes (NB):

NB is a statistical and probabilistic classification technique that is one of the most successful. It calculates the probability of belonging to a class using the data we use to decide to which class a sample belongs [32]. This is done by assuming that the influence of an attribute value on a class is independent of the impact of other attribute values [33].

3.4.1.4 Logistic Regression (LR):

LR is a type of generalized linear regression model widely used to predict the probability of occurrence of an event [34] . In logistic regression models, the dependent variable is always in categorical form and has two or more levels, the independent variables can be in numerical or categorical form [22]. In this study, we used this supervised classification model to predict effective combination therapies that reduce the risk of metastatic relapse in breast cancer patients.

3.4.1.5 Artificial neural network (ANN):

The artificial neural network, also known as a connectionist system, is a computer system loosely based on the human brain. The algorithm consists of a number of highly interconnected nodes organized in layers, which allows the information processing architecture to recognize complex, non-linear patterns between input data and output variables [35].

3.4.1.6 Random forest (RF):

Breiman introduced Random Forests (RF) as a tree-based ensemble learning approach for classification and regression in 2001 [36]. It has been frequently used in the healthcare field due to its simple structure and superior performance compared to other machine learning approaches.

In the next paragraph, we will evaluate the efficiency of all classifiers in terms of time to build the model, correctly classified instances, incorrectly classified instances and accuracy.

3.5 Performance indicators

Each model was evaluated on the metrics of accuracy, sensitivity, specificity, precision, recall curve, and finally, area under the receiver operating characteristic curve (AUC).

3.5.1 Confusion Matrix

A confusion matrix is commonly used to visualize the performance of a classification algorithm. Figure 4 shows the confusion matrix for a multi-class model with N classes [37]. Observations on correct and incorrect classifications are collected in the confusion matrix C (Cij), where Cij represents the frequency with which class i is identified as class j. In general, the confusion matrix provides four types of classification results with respect to a classification target k:

- ✓ True positive (TP): the prediction of the positive class is correct ($c_{K,K}$)
- ✓ True negative (TN): correct prediction of the negative class $\sum_{i,j\in N\backslash\{k\}} c_{ij}$
- ✓ False positive (FP): incorrect prediction of the positive class $\sum_{i\in N\backslash\{k\}} c_{ik}$
- ✓ False negative (FN): incorrect prediction of the negative class $\sum_{i\in N\backslash\{k\}} c_{ki}$

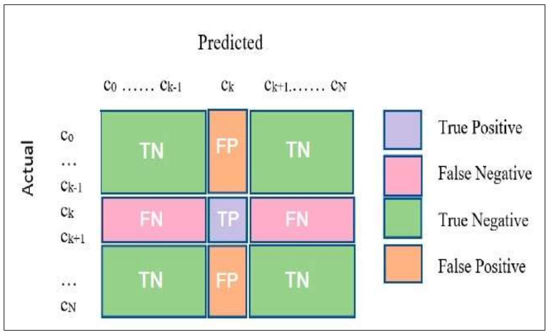


Fig.4 Confusion Matrix For Multi-Class Classification

3.5.2 Classification report:

A classification report is used to summarize the performance of the classification model.

Accuracy is the proportion of the total number of correct predictions. It is defined as the total number of positive instances of the model divided by the total number of instances. The accuracy parameter provides the percentage of correctly classified instances. The accuracy of the model is defined as:

$$Overall\ Accuracy = \frac{\sum_{i=1}^{N} c_{i,i}}{\sum_{i=1}^{N}\sum_{j=1}^{N} c_{i,j}} \quad (1)$$

Precision (2) is the ratio of true positives to all positives. For our problem statement, this parameter is used to determine the degree of the attribute to correctly classify the combination of effective adjuvant treatments, is defined as:

$$Precision_{class} = \frac{TP_{class}}{TP_{class} + FP_{class}} \quad (2)$$

The true negative rate (**Specificity**) is defined by equation (3). The false positive rate is the proportion of negative data points that are correctly considered negative, out of all negative data points.

$$Specificity_{\ class} = \frac{TN_{class}}{FP_{class} + TN_{class}} \quad (3)$$

The recall (**Sensitivity**) is the true positive rate defined by equation (4). This rate is the proportion of positive data points that are correctly considered as positive, on all positive data points.

$$Recall_{\ class} = \frac{TP_{class}}{TP_{class} + FN_{class}} \quad (4)$$

Sensitivity and **specificity** are also called quality parameters and used to define the quality of the predicted class. To determine the quality of the medical diagnostic model, three parameters are basically used; these three parameters are accuracy, sensitivity and specificity.

F1-Score: This is a harmonic average metric of accuracy and Recall. Although the F1-Score is not as intuitive as Precision, it is useful for measuring the accuracy and robustness of the classifier[38].

The Roc and AUC curve:

A receiver operating characteristic (ROC) curve is a curve that plots the true positive rate (sensitivity) against the false positive rate (1 - specificity) as the decision threshold changes [31]. The area under the curve (AUC) can be interpreted as a measure of the probability that the model will rank a positive random example over a negative random example. Its values range from 0 to 1. A model with 100% wrong predictions has an AUC of 0. If all its predictions are correct, its AUC is 1.The comparison of the performance of the learning algorithms, discussed in the next section, is based on these indicators (Accuracy; Precision; Specificity; Recall; AUC).

4. RESULTS AND DISCUSSION

4.1 Analysis of Result

In this study, the quality of the multinomial logistic regression model is evaluated by the classification methods and the confusion matrix. We used the variables (age at diagnosis of breast cancer, postmenopausal, primary tumor size (TS), histological grade of the tumor, number of involved axillary lymph nodes, cellular marker of proliferation (Ki67), type of surgery), the characteristics ER, PR and HER2 are also included in our breast cancer registry dataset, to achieve better performance (in terms of ROC surface and accuracy) without sacrificing accuracy. All results below are 10-fold cross-validation results, each representing the optimal outcome of this method, the results of these classification algorithms are shown in Table 6 below.

Table 5: The Multi-Class Confusion Matrix Of The Classification Models Used

A : Classifier (Random Forest)

		Predicted				
		11105	21010	22011	32012	Σ
Current	11105	73	0	0	0	73
	2010	0	47	1	2	50
	22011	0	0	84	13	97
	32012	0	1	10	70	81
Σ		73	48	95	85	301

B : Classifier (Naive Bayes)

		Predicted				
		11105	21010	22011	32012	∑
Current	11105	71	0	2	0	73
	21010	0	47	1	2	50
	22011	0	1	79	17	97
	32012	0	1	10	70	81
∑		71	49	92	89	301

C : Classifier (ANN)

		Predicted				
		11105	21010	22011	32012	∑
Current	11105	73	0	0	0	73
	21010	0	47	1	2	50
	22011	0	0	80	17	97
	32012	0	1	14	66	81
∑		73	48	95	85	301

D : Classifier (Decision Tree)

		Predicted				
		11105	21010	22011	32012	∑
Current	11105	73	0	0	0	73
	21010	0	46	2	2	50
	22011	0	0	84	13	97
	32012	0	4	18	59	81
∑		73	50	104	74	301

E : Classifier (Logistic regression)

		Predicted				
		11105	21010	22011	32012	∑
Current	11105	73	0	0	0	73
	21010	0	47	1	2	50
	22011	0	0	81	16	97
	32012	0	1	22	58	81
∑		73	48	104	76	301

F : Classifier (SVM)

		Predicted				
		11105	*21010*	*22011*	*32012*	∑
Current	*11105*	*73*	*0*	*0*	*0*	*73*
	21010	*0*	*47*	*1*	*2*	*50*
	22011	*0*	*0*	*75*	*22*	*97*
	32012	*0*	*1*	*22*	*58*	*81*
∑		*73*	*48*	*98*	*82*	*301*

The results showed that the Random Forest classifier was more accurate in predicting good adjuvant therapy combinations and the highest falsely predictive number was 13 for adjuvant therapy combination 32012, followed successively by Naive Bayes, Neural Network, Decision Tree, Logistic Regression and SVM. On the other hand, we find that the SVM is the highest in terms of false predictions with the highest number (22 false predictions), for the combined adjuvant therapy code 32012.

4.2 Performance Evaluation

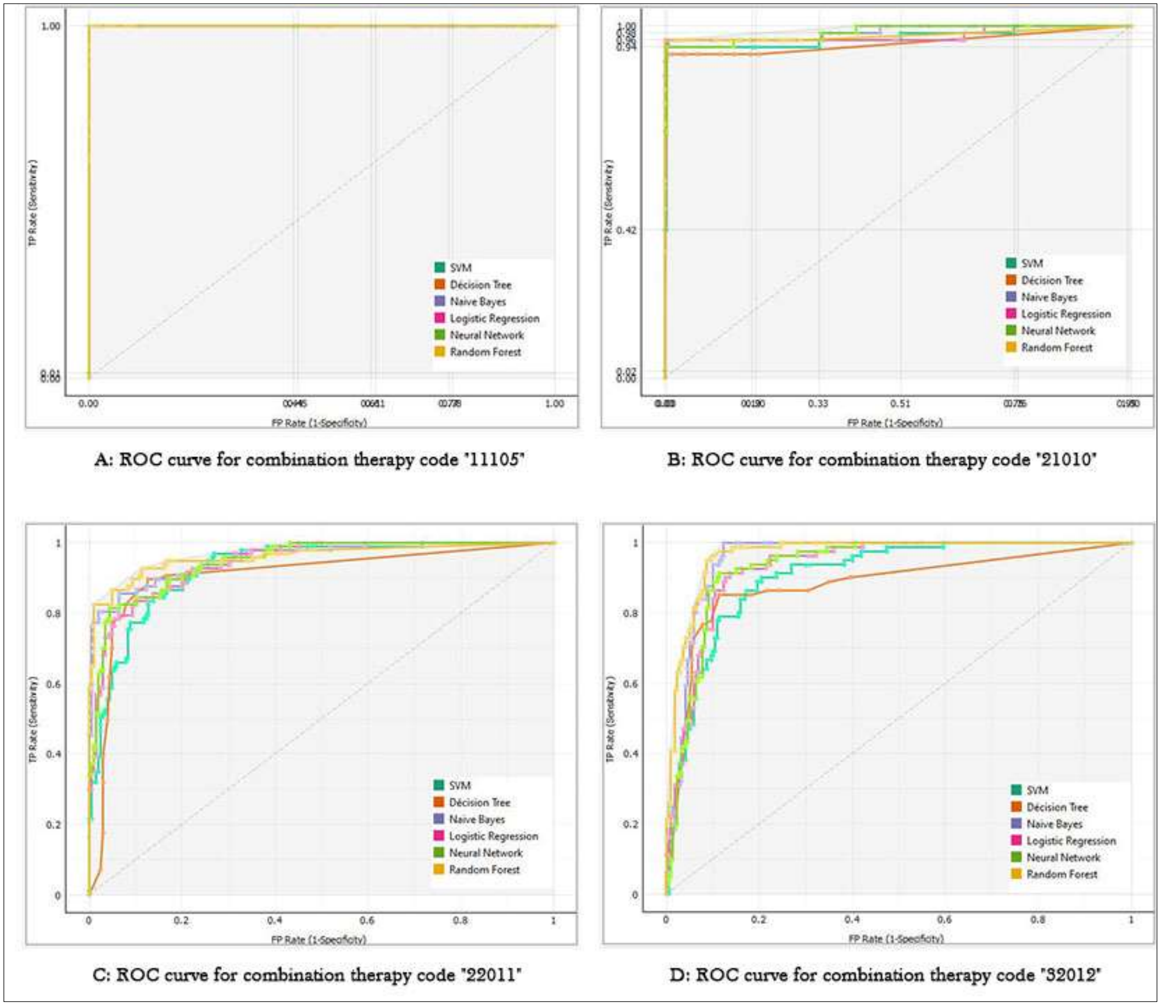
Classification measures were calculated to compare the performance of the six algorithms. Table 7 shows that the Random Forest algorithm obtained the best results in terms of accuracy (91%), sensitivity (91%), specificity (91.2%) and f1 measure (91.1%). Considering AUC, Random Forest also had the highest specificity (97.4%).

Table 6: Evaluation Of The Different Machine Learning Algorithms Used

Model	AUC	CA	F1	Precision	Recall
RF	0.974	0.910	0.911	0.912	0.910
NB	0.972	0.887	0.888	0.890	0.887
NN	0.963	0.884	0.884	0.886	0.884
DT	0.931	0.870	0.870	0.870	0.870
LR	0.960	0.860	0.860	0.862	0.860
SVM	0.948	0.841	0.841	0.842	0.841

Roc and AUC curve:

All machine learning classifiers give an accuracy level of more than 84% for the classification of the combination of adjuvant therapies, which reduces the risk of recurrence in breast cancer patients. This indicates that the performance of these classification techniques is excellent for predicting the combination of adjuvant therapies. It can be inferred that it is very important to know the receiver operating characteristic (ROC) curve, which is based on the true positive rate (TPR) and false positive rate (FPR) of these classification results[39]. According to the ROC curve (see Figure 5), Random Forest achieved the highest AUC (area under the curve) for ROC in the following adjuvant therapy combination codes (11105 - 21010 - 22011- 32012).



A: ROC curve for combination therapy code "11105"

B: ROC curve for combination therapy code "21010"

C: ROC curve for combination therapy code "22011"

D: ROC curve for combination therapy code "32012"

Fig.5 ROC Curve For The Four Predicted Variables That Signify The Therapeutic Combinations Used In This Study

We observe in Fig5 that the ROC and AUC curves obtained from the test data set show the areas under the ROC curves with similar patterns in the upper left corner. This means that the classifiers correctly predict the therapeutic combination code 11105, followed successively by 21010, 22011, and 32012 presented in the ROC curves (B, C, D).

Although machine learning showed good results in predicting the combination of effective adjuvant treatments for breast cancer, this study has some limitations. Some patients were excluded due to lack of data, which may lead to selection bias. In addition, because of the retrospective data, our study failed to refine the prediction of the optimal combination of adjuvant therapies in certain subgroups of the postoperative breast cancer population, such as patients with breast cancer associated with other malignancies and patients with breast cancer with other specific medical histories, which may lead to some applicability of the study results. Further prospective studies on this aspect are needed in the future.

5. CONCLUSION

In this paper we proposed a new supervised multinomial logistic regression model to predict optimal adjuvant treatments for breast cancer patients that reduce metastatic relapse. We also presented a comparative study of different multi-class machine learning algorithms (Logistic Regression (LR), Naive Bayes (NB), Random Forest (RF), Decision Tree (DT), Support Vector Machine (SVM), and Artificial Neural Network (ANN). The result obtained shows that the Random Forest classifier gives a better result in terms of accuracy and low error rate in the prediction of adjuvant therapy treatment protocols.

The classification used in the proposed system is a multi-class based on the four most recommended adjuvant treatment protocols in the Meknes oncology center in Morocco. The experimental results of the evaluation tests show that the proposed model achieves the research goals by drawing attention to a new method that determines the best accurate results after studying the data mining techniques in the previous works. Also, allowing the classification of breast cancer treatment strategies. In addition to determining whether the new treatment will be beneficial or not for the patient, which reduces metastatic relapses, is expensive and avoids unnecessary procedures. concluding that the proposed model has the potential to improve care, save lives, reduce costs and make more informed decisions. However, some limitations could threaten the proper implementation and prevent the benefits of applying the proposed model, such as data accessibility, data collection and availability, data sample size, complexity of the analysis, which increases the calculation time.

REFERENCES:

[1] F. Bray, J. Ferlay, I. Soerjomataram, R. L. Siegel, L. A. Torre, and A. Jemal, “Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries,” *CA: A Cancer Journal for Clinicians*, vol. 68, no. 6, Art. no. 6, Nov. 2018, doi: 10.3322/caac.21492.

[2] C. emile, “Prise en charge du cancer du sein,” *Option/Bio*, vol. 28, no. 565–566, Art. no. 565–566, Jul. 2017, doi: 10.1016/S0992-5945(17)30176-9.

[3] G. Jerusalem *et al.*, “Cancer du sein : de la thérapie ciblée à la médecine personnalisée,” *Rev Med Liège*, p. 8.

[4] S. Chopra and E. L. Davies, “Breast cancer,” *Medicine*, vol. 48, no. 2, pp. 113–118, Feb. 2020, doi: 10.1016/j.mpmed.2019.11.009.

[5] S. Hubert and J.-P. Abastado, “Les étapes précoces du processus métastatique,” *Med Sci (Paris)*, vol. 30, no. 4, pp. 378–384, Apr. 2014, doi: 10.1051/medsci/20143004010.

[6] M. Scimeca *et al.*, “Novel insights into breast cancer progression and metastasis: A multidisciplinary opportunity to transition from biology to clinical oncology,” *Biochimica et Biophysica Acta (BBA) - Reviews on Cancer*, vol. 1872, no. 1, Art. no. 1, Aug. 2019, doi: 10.1016/j.bbcan.2019.07.002.

[7] M. Scimeca *et al.*, “Novel insights into breast cancer progression and metastasis: A multidisciplinary opportunity to transition from biology to clinical oncology,” *Biochimica et Biophysica Acta (BBA) - Reviews on Cancer*, vol. 1872, no. 1, Art. no. 1, Aug. 2019, doi: 10.1016/j.bbcan.2019.07.002.

[8] J. Hayward, “Mining Oncology Data: Knowledge Discovery in Clinical Performance of Cancer Patients,” p. 270.

[9] J. S. Akosa, “Application of Data Mining Techniques in Improving Breast Cancer Diagnosis,” p. 10.

[10] B. R. Andjelkovic Cirkovic, “Machine learning approach for breast cancer prognosis prediction,” in *Computational Modeling in Bioengineering and Bioinformatics*, Elsevier, 2020, pp. 41–68. doi: 10.1016/B978-0-12-819583-3.00002-3.

[11] Z. Tasnim, “Classification of Breast Cancer Cell Images using Multiple Convolution Neural Network Architectures,” *International Journal of Advanced Computer Science and Applications*, vol. 12, no. 9, p. 8, 2021.

[12] A. M. Sayed, “Machine Learning Augmented Breast Tumors Classification using Magnetic Resonance Imaging Histograms,” *IJACSA*, vol. 12, no. 12, 2021, doi: 10.14569/IJACSA.2021.0121201.

[13] A. Mailliez, C. Decanter, and J. Bonneterre, “Chimiothérapie adjuvante de cancer du sein et fertilité : estimation de l’impact, options de préservation et place de l’oncologue,” *Bulletin du Cancer*, vol. 98, no. 7, Art. no. 7, Jul. 2011, doi: 10.1684/bdc.2011.1391.

[14] K. S. Asgeirsson *et al.*, “Serum epidermal growth factor receptor and HER2 expression in primary and metastatic breast cancer patients,” *Breast Cancer Res*, vol. 9, no. 6, Art. no. 6, Dec. 2007, doi: 10.1186/bcr1788.

[15] D. J.-P. Zurcher and A. Stravodimou, “Hormonothérapie dans le cancer du sein invasif, update 2016,” *REVUE MÉDICALE SUISSE*, p. 4, 2016.

[16] M. Guilabert *et al.*, "A Web-Based Self-assessment Model for Evaluating Multidisciplinary Cancer Teams in Spain: Development and Validation Pilot Study," *J Med Internet Res*, vol. 24, no. 3, p. e29063, Mar. 2022, doi: 10.2196/29063.

[17] N. M. Swelam, A. E. Khedr, and H. Auda, "BREAST CANCER DIAGNOSIS AND PROGNOSIS USING STACKING ENSEMBLE TECHNIQUE," . *Vol.*, no. 14, p. 14, 2022.

[18] M. A. Elsadig, "ENSEMBLE CLASSIFIER FOR BREAST CANCER DETECTION," . *Vol.*, no. 10, p. 10, 2022.

[19] G. C. Wishart *et al.*, "RPeRseaErcDh aIrCticlTe : a new UK prognostic model that predicts survival following surgery for invasive breast cancer," *Breast Cancer Research*, p. 10, 2010.

[20] S. P. Somashekhar *et al.*, "Watson for Oncology and breast cancer treatment recommendations: agreement with an expert multidisciplinary tumor board," *Annals of Oncology*, vol. 29, no. 2, pp. 418–423, Feb. 2018, doi: 10.1093/annonc/mdx781.

[21] H. Akhmouch, H. Bouanani, G. Dias, and J. G. Moreno, "Stratégie Multitâche pour la Classification Multiclasse," p. 10.

[22] E. Bisong, "Logistic Regression," in *Building Machine Learning and Deep Learning Models on Google Cloud Platform*, Berkeley, CA: Apress, 2019, pp. 243–250. doi: 10.1007/978-1-4842-4470-8_20.

[23] M. Lango and J. Stefanowski, "Multi-class and feature selection extensions of Roughly Balanced Bagging for imbalanced data," *J Intell Inf Syst*, vol. 50, no. 1, pp. 97–127, Feb. 2018, doi: 10.1007/s10844-017-0446-7.

[24] T. J. Cole, "Applied logistic regression. D. W. Hosmer and S. Lemeshow, Wiley, New York, 1989. No. of pages: xiii + 307. Price: £36.00," *Statist. Med.*, vol. 10, no. 7, pp. 1162–1163, Jul. 1991, doi: 10.1002/sim.4780100718.

[25] D. G. Kleinbaum and M. Klein, "Ordinal Logistic Regression," in *Logistic Regression*, New York, NY: Springer New York, 2010, pp. 463–488. doi: 10.1007/978-1-4419-1742-3_13.

[26] M. Ertel and S. Amali, "'Artificial Intelligence (AI) in oncology: Predicting Treatment Response in Women with Breast Cancer'. (2021). 1st International Meeting on science at the service of health in the context of public-private partnership. May 27 - 28, Meknes, Morocco.," 2021.

[27] "Performance of Support Vector Machine Kernels (SVM-K) on Breast Cancer (BC) Dataset," *ijrte*, vol. 8, no. 2S7, Art. no. 2S7, Sep. 2019, doi: 10.35940/ijrte.B1076.0782S719.

[28] N. E. Khalifa, G. Manogaran, M. H. N. Taha, and M. Loey, "THE CLASSIFICATION OF POSSIBLE CORONAVIRUS TREATMENTS ON A SINGLE HUMAN CELL USING DEEP LEARNING AND MACHINE LEARNING APPROACHES," . *Vol.*, no. 21, p. 12, 2021.

[29] R. M. Farag, M. A. El-Dosuky, and M. Z. Rashad, "DNA SEQUENCE ANALYSIS FOR DISEASE PREDICTION AND TREATMENT BASED ON MACHINE LEARNING," . *Vol.*, no. 24, p. 14, 2021.

[30] Q. Lin and J. Son, "AN IN-SHIP LOCALIZATION ALGORITHM FOR CLOSE CONTACT IDENTIFICATION," . *Vol.*, no. 6, p. 12, 2022.

[31] "Syarif et al. - 2002 - Study on multi-stage logistic chain network a spa.pdf."

[32] I. Kononenko, "INDUCTIVE AND BAYESIAN LEARNING IN MEDICAL DIAGNOSIS," *Applied Artificial Intelligence*, vol. 7, no. 4, pp. 317–337, Oct. 1993, doi: 10.1080/08839519308949993.

[33] A. Jamain and D. J. Hand, "The Naive Bayes Mystery: A classification detective story," *Pattern Recognition Letters*, vol. 26, no. 11, pp. 1752–1760, Aug. 2005, doi: 10.1016/j.patrec.2005.02.001.

[34] M. Kim, "Two-stage logistic regression model," *Expert Systems with Applications*, vol. 36, no. 3, pp. 6727–6734, Apr. 2009, doi: 10.1016/j.eswa.2008.08.063.

[35] P. J. Drew and J. R. T. Monson, "Artificial neural networks," vol. 127, no. 1, p. 9, 2000.

[36] L. Breiman, "Random Forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, Oct. 2001, doi: 10.1023/A:1010933404324.

[37] F. Krüger, "Activity, context, and plan recognition with computational causal behavior models," 2018, doi: 10.18453/ROSDOK_ID00002015.

[38] H. B. Nembhard, "Statistical Process Adjustment Methods for Quality Control," *Journal of the American Statistical Association*, vol. 99, no. 466, pp. 567–568, Jun. 2004, doi: 10.1198/jasa.2004.s340.

[39] "Quinlan et al. - A Comparative Analysis of Classification Techniques.pdf."